

Predicting Small Group Accretion in Social Networks: A topology based incremental approach

Ankit Sharma, Rui Kuang and Jaideep Srivastava
Department of Computer Science and Engineering
University of Minnesota, Minneapolis, MN 55455
{ankit,kuang,srivasta}@cs.umn.edu

Xiaodong Feng
University of Electronic Science
and Technology of China, China
fxdong88@gmail.com

Kartik Singhal
LNM Institute of Information
Technology, Jaipur, India
singh559@umn.edu

Abstract—Small Group evolution has been of central importance in social sciences and also in the industry for understanding dynamics of team formation. While most of research works studying groups deal at a macro level with evolution of arbitrary size communities, in this paper we restrict ourselves to studying evolution of small group (size ≤ 20) which is governed by contrasting sociological phenomenon. Given a previous history of group collaboration between a set of actors, we address the problem of predicting likely future group collaborations. Unfortunately, predicting groups requires choosing from $\binom{n}{r}$ possibilities (where r is group size and n is total number of actors), which becomes computationally intractable as group size increases. However, our statistical analysis of a real world dataset has shown that two processes: an external actor joining an existing group (*incremental accretion* (IA)) or collaborating with a subset of actors of an existing group (*subgroup accretion* (SA)), are largely responsible for future group formation. This helps to drastically reduce the $\binom{n}{r}$ possibilities. We therefore, model the attachment of a group for different actors outside this group. In this paper, we have built three topology based prediction models to study these phenomena. The performance of these models is evaluated using extensive experiments over DBLP dataset. Our prediction results shows that the proposed models are significantly useful for future group predictions both for IA and SA.

Keywords—Social Networks, Higher Order Link Prediction, Group Evolution, Hypergraphs, Hypergraph Evolution

I. INTRODUCTION

Study of small groups has been an important endeavor in a large number of disciplines like psychology, sociology, communication and information science for past 50 years [1]. Advent of globalization has lead to changing nature of groups or teams in industry leading to increasing interest in studying their dynamics [2]. Large part of the research in the field of Organization Science is dedicated to study effective ways to build teams by combining employee expertise [3]. With the rising number of large interdisciplinary scientific teams [4], understanding drivers affecting their success is of key importance for science funding agencies while selecting team of scientists [5]. Other real life applications include building emergency response teams for natural disasters management, automation of team selection for military operations and self-organizing open-software teams [6]. While such studies are important, the ever increasing availability of online “group” interaction data for example, social networking sites like Facebook or Twitter, group communication tools like Skype, Google Hangout, Google Docs, Massive Online multi-player games (MMOGs) such as World of Warcraft, etc., makes such studies even more realistic. In scientific research, such dataset

have been used to study group dynamics for benefit of both industry and academia [7].

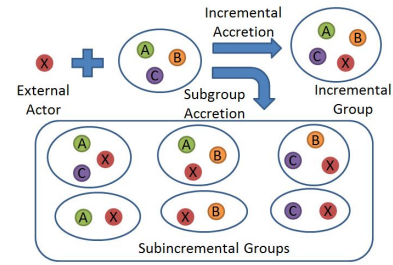


Figure 1. Example illustrating *incremental accretion* and *subgroup accretion* processes. Set {A,B,C} is a group and X is an external actor.

Several studies in the past have analyzed the usefulness of various features used for membership prediction in communities [8][9][10][11]. Also, some works have focused on group membership dynamics by simulating how individual join or leave a group [12][13][14], while such studies have not developed prediction models. Objective of our work is to focus on the task of actual future group prediction in contrast to feature analysis or simulation based studies. Moreover, most past works on group evolution in social networks primarily deal with evolution of arbitrary size communities or groups [15][16][17]. These sizes are usually large and the boundaries of community depend on the definition of membership employed [18]. In this work we are interested in well defined small groups (size ≤ 20) like research collaborations or teams [19] also called as *Bona fide groups* [20] in sociology. Also these small groups are self assembling [7] where members leave or join groups autonomously and the motivation or theories for group formation is much different from the large communities [1]. Moreover, these groups are connected by the social network of actors, resulting in group ties or *network of groups* [21] (see Figure 2), which plays central role in group formation process. We focus on group accretion which is a subset of the group evolution. Group accretion is the process of size increment in groups by addition of more members. More specifically we define two subproblems: *incremental accretion* and *subgroup accretion*. In the first problem, given a group of size= x , we predict the likelihood of one more member being absorbed in it, to yield a size= $(x + 1)$ incremental group (Figure 1). The subgroup accretion is the problem of incremental accretion on all the $(2^x - 2)$ subgroups of a given group to yield prediction scores of $(2^x - 2)$ new incremental groups (Figure 1). Intuition behind choice of these problems is that, given a past history of group collaborations, a large

percentage of groups in future are formed through these two processes. Our aim finally is to build models that predict future groups that are likely to be formed using these two mechanisms. This work therefore, is an initial step towards a more general higher order group prediction problem.

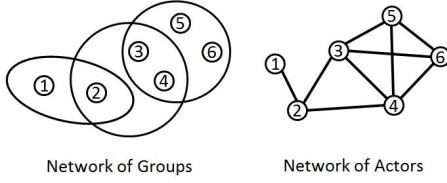


Figure 2. Example illustrating *network of groups* (left) and the corresponding *network of actors* (right) where $\{1,2,3,4,5,6\}$ are the actors

In this paper we assume that the groups are not isolated but rather interact with each other through *network of groups* and group members make individual decisions to collaborate or not (*self assembly*). We therefore, model the attachment of a group (*group*) to an actor (*node*) outside it. Topology based dyadic link prediction (DLP) methods have proven successful in capturing *node to node* attachment. Guided by both, DLP methods and sociology theories of small groups we have proposed three different methods. First, method is an extension of the popular path enumerating Katz method [22]. Second, is a network alignment based supervised method. With a philosophy similar to Katz, it captures the inter-group communication cycles, which are theoretically hypothesized more appropriate, rather than paths. Lastly, we also propose a label propagation based method where the random walks are guided by the network of groups (a hypergraph [23]). Based on the extensive set of experiments it turns out our models are able to effectively predict future groups formed by both IA and SA processes.

Now we summarize the key points of this research paper:

- We highlight an important observation that a large chunk of future groups are formed by two different accretion processes from the past groups.
- We have focused on the evolution of small groups and defined new problems for small group accretion.
- We have developed three topology based methods, guided by social theories, to address these problems in a novel manner.
- We have therefore, presented an overall new incremental approach to address the less addressed and intriguing problem of higher order link prediction.

II. RELATED WORKS

Social Group Evolution There is a vast body of literature regarding community evolution [18] and more general network evolution [24] in social networks. Definition of community in all these works varies drastically producing from small groups to size as large as hundreds or thousands. Therefore, they take a macroscopic view while answering the questions of how to detect communities and how they evolve over time. Though, recently there have been some works which zoom in and try to understand the evolution from the perspective of individual actors and their relationship with other actor group.

Among them the first category of works focus on building models which can simulate the group formation tasks like leaving, joining or switching between groups [12][13][14]. Alvari et. al. [25] provide a game theoretic community/group detection model where the actors in the social network are rational agents performing these tasks while maximizing their utility. Same authors extend their work for evolutionary setting [26] and apply to MMOGs [14]. MMOG guild formation is also studied as stochastic processes in both network [13] and network-less settings [12]. Our work, rather than simulation, focuses on predicting the exact groups that might occur in future.

Second category deals with analysis of various characteristics of groups (like diversity, cohesion, stability, type, etc.) and their correlation with different factors (size, member properties, etc.) [27][15].

Third category of papers are devoted to extracting features of different kinds like network, actor, group or communication content and pose the problem of group membership prediction as a classification task. For instance, Patil et al. have done feature extraction for group attrition [16], group stability [8] as well as group destruction [9]. It differs from our work as they focus on understanding the importance of various features rather than prediction of groups. In a series of papers from Sharara et al. have stressed on the idea of loyalty (affinity) of actor towards different groups and its longitudinal changes [17]. This idea has been extended for deducing important actors by using group semantics (like diversity) [10] and applied to analysis of guilds in MOMGs [11]. Both, Patil et al. and Sharara et al. model actor's attachment or loyalty for different groups whereas we model the opposite: tendency of a group to absorb different actors. Moreover, both of them primarily deal with large communities like conferences in DBLP data. We are specifically focused on predicting small cohesive groups or teams (like a small group of researcher working on a publication in DBLP) where different social phenomenon are at play. We rather treat the problem as extension of DLP to higher order (group) prediction.

Social Sciences Naturally occurring small groups are called *Bona Fide Groups* [20], which are the focus of this paper. A good reference for small group theories can be found in Poole et al. [1]. Plethora of research deals with application of small group dynamics for understanding teams [3]. Network perspective of teams has been a new development in the past decade. Various studies like Oh et. al. [28][5] stress the importance of network structure in determining team performance. More recent research has focused on self-assembling teams in which members autonomously leave or join teams [7][5].

Link Prediction Due to space constraint we point out some key surveys and papers for DLP. An overview of link prediction in general complex network is by Lu et al. [29] and more specifically for social network we refer to Hasan et al. [30]. We have generalized topology based methods like Katz (1953) [22], proven successful for social networks [31], for higher order links. For network alignment based link prediction we refer Flannick et al [32] and Xie et al. [33].

The rest of the paper is as follows: Section III describes the problem statement, Section IV describes the topology based methods, in Section V the experiments conducted are described and results are discussed, followed by conclusion in section VI.

III. PROBLEMS AND PRELIMINARIES

A. Problem Definition

In this paper we consider the scenario where we have a set of individuals or social *actors*. These *actors* self assemble themselves into *groups* to perform tasks at hand or gather for an event. A *group* therefore, is a subset of all the *actors*. Membership of a *group* can change over time. An *actor* can leave or join a *group*, resulting in changes in *group* membership. When two *actors* work or gather together in the same *group* they develop social tie. These social ties therefore, become the edges in the social *network of actors* (NOA). Moreover, the *actors* that are the part of multiple groups act as ties between groups resulting in a *network of groups* (NOG) (as we had mentioned in introduction). Given a past history of *groups* formed our problem is to predict *groups* that are likely to form in future by two different evolutionary processes described as follows. Given a *group*, it can absorb an *actor* outside this *group* to form a new group in future. This process is called *incremental accretion* (IA) and the group formed by this process is called *incremental group* (IG) (Figure 1). In the second process, rather than all the members of a given *group*, only a subset of them absorb an actor outside the group to form another *group*. This process is *subgroup accretion* (SA) and the corresponding group formed is called *subincremental group* (SG) (Figure 1). Note that it is possible that SG and/or IG might have been previously observed or not observed in history. We therefore restrict ourselves to predict only the IG and SG type groups formed by IA and SA processes respectively by assigning prediction scores to them.

An example of such a scenario is collaborations among authors to work on publication. As authors write papers they develop social relations with each other. As authors work in multiple research collaborations they become intermediates between these different research collaborations. Related example can be open-source software development teams.

B. Problem Statement

We have a set of n actors $V = \{v_1, v_2, \dots, v_n\}$. A subset of these *actors* form a *group*. We have a collection of m such groups observed in past, denoted by $G = \{g_1, g_2, \dots, g_m\}$ where $g_i \subseteq V$ represents the i^{th} group. Cardinality $c_i = |g_i|$ of a group is the number of actors part of it. We have two networks. First, NOG is a hypergraph [23] represented as a set $N_g = (V, G)$ with G as the hyperedges over the vertex set V . We also have an incidence matrix \mathbf{H} for N_g of size $(|V| \times |G|)$ with elements defined as $\mathbf{H}(v, g) = 1$ if $v \in g$ else 0. Second, NOA is a graph, $N_a = (V, E)$ where $E = \{e_1, \dots, e_w\}$ are the dyadic edges defined over vertex set V . Adjacency matrix \mathbf{A} of size $(|V| \times |V|)$ for N_a has elements $\mathbf{A}(p, q) = 1$ for $(p, q) \in E$ such that $\exists i, \{p, q\} \subseteq g_i$ else 0.

In IA, a *group* $g_i \in G$ can absorb an actor $a \in \{V - g_i\}$ to produce $g_i^a = \{g_i \cup a\}$. Let $g_i^{in} = \{g_i^a\}_{a \in \{V - g_i\}}$ be the set of all the IGs for i^{th} group. Our aim therefore, in IA problem is to predict a score to each of the IGs in set $g_i^{in} \forall i \in \{1, 2, \dots, m\}$.

Considering the second case of SA problem. We define a proper subgroup of g_i as $s_i \subset g_i$ where $s_i \neq \phi$. A subgroup s_i can absorb an actor $a \in \{V - s_i\}$ to produce $s_i^a = \{s_i \cup a\}$. Let $g_i^{sa} = \{s_i^a\}_{a \in \{V - s_i\}}$ be the set of all the SGs for i^{th} group. Our aim therefore, in SA problem is to predict a score to each of the SGs in set $g_i^{sa} \forall i \in \{1, 2, \dots, m\}$.

IV. METHODS

In this section we describe three methods to solve the problems described in the previous section. Each method models the affinity of a given *group* towards an *actor* outside the group in different ways. As pointed out earlier, these methods are inspired from the existing dyadic link prediction (DLP) techniques as well as sociology theories. Success of topology based DLP methods encouraged us to focus on topology derived methods. First, method is a generalization of unsupervised path counting based DLP methods to predict the IGs and SGs. The second approach, is a semi-supervised learning based on network alignment algorithms [32]. It captures the cycles that pass through both the given group and the rest of graph. The third approach, uses a semi-supervised hypergraph label propagation approach. Each of these methods provide a score $\mathbf{S}(i, j)$ between i^{th} group and j^{th} actor, representing the similarity or affinity between them.

A. Generalized Katz Score (GKS)

Among the similarity score based methods in DLP, methods based on counting ensemble of paths have been most successful. More specifically Katz (1953) [22] measure has been shown to outperform all other similarity scores [31]. Given the NOA graph N_a the Katz Score (KS) between any two nodes i and j is defined as:

$$\mathbf{K}(i, j) = \sum_{l=1}^{\infty} \beta^l |\text{paths}_{i,j}^{(l)}| \quad (1)$$

where \mathbf{K} is $(|V| \times |V|)$ size matrix containing KS for different pair of vertices, $|\text{paths}_{i,j}^{(l)}|$ is number of l length paths between the nodes i and j , and $\beta \in (0, 1)$ is parameter that controls the extent to which long paths are penalized. From a sociology perspective number of path between any two nodes capture their social proximity. For example two scientists, in a co-authorship network, who are close to each other in this network, should have many common colleagues or are in similar social circles. Therefore, they are more likely to collaborate. Also smaller length paths reflect greater proximity and are more important, hence, are less penalized or contribute more. In matrix terms it is calculated as:

$$\mathbf{K} = \sum_{l=1}^{\infty} \beta^l \mathbf{A}^l \quad (2)$$

where \mathbf{A} is the adjacency matrix for N_a . Success of this subtle KS method has encouraged us to generalize it to capture the affinity of a *group* for an *actor* outside it. We therefore, define the Generalized Katz Score (GKS) between i^{th} group and j^{th} actor node as follows:

$$\mathbf{S}(i, j) = \frac{1}{c} \sum_{p \in g_i} \sum_{l=1}^{\infty} \beta^l |\text{paths}_{p,j}^{(l)}| = \frac{1}{c} \sum_{p \in g_i} \mathbf{K}(p, j) \quad (3)$$

where g_i is the i^{th} group of cardinality c . This score is the average proximity, measured as KS, of different *actors* within the group to a given external *actor*. For higher order groups the proportion of the group members close to an external individual is more relevant. Therefore, taking average is intuitive, as it tells that on an average how close is this external individual to the given group. It captures the chances he might get absorbed in this group by taking into account the size of group.

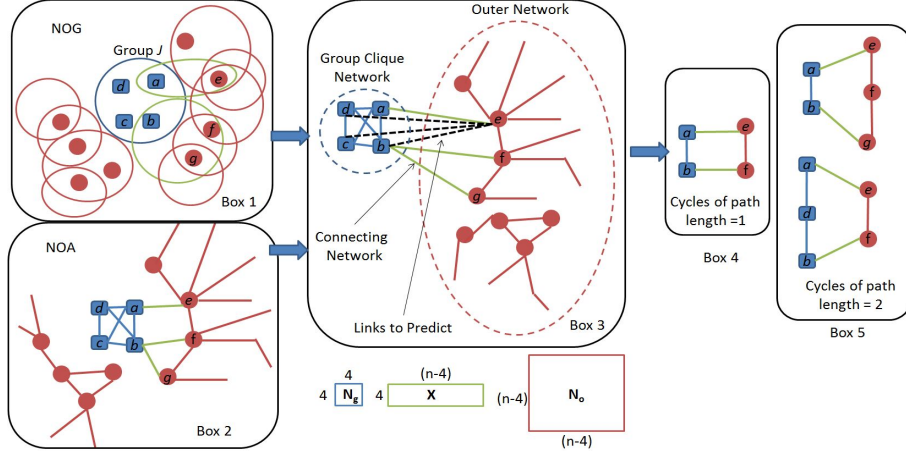


Figure 3. Example illustrating different networks used in BRWS for a sample group J consisting of actors $\{a, b, c, d\}$. Box 1 and Box 2 shows the NOG and NOA around Group J actors. In Box 3, we have the 3 network: group clique network (blue), the outer network (red) and bipartite inter network (green). Adjacency matrices used in BRWS are at bottom. Box 4 and Box 5 are example of some cycles of different maximum path length involving actor e . These example cycles (along with other cycles not shown) are used in BRWS to calculate scored of the black dotted edges from group members to actor e . Note for top cycle in box 5 has path length of 2 over outer network and 1 over clique network. However, these lengths are vice versa for the bottom cycle.

B. Bi-random Walk Score (BRWS)

GKS makes use of the communication paths to measure the affinity of the each group members to a person outside the group separately. In this section, we model the scenario when outside individual is known by multiple members of the group. Consider the group J (blue color) in Figure 3 with four actors $\{a, b, c, d\}$ in it. Let us take the external actor e for whom we wish to observe affinity with J . We can observe that e has a direct link only with the member a . But, b has a direct link with a neighbor f of e . Though, there is no direct relation between b and e , but they have an indirect link through actors both internal (a) as well as external (f) to the group. To quantify the affinity between b and e , quantification process should be guided by both these internal and external links. We model this intuition in a holistic way by capturing the cycles like $\{a \rightarrow b \rightarrow f \rightarrow e \rightarrow a\}$. Instead of simply counting, we use these cycles to learn the affinity of group member to an external actor. We cast this scenario as an alignment problem [32] where the nodes within group have to be aligned with external nodes. One of the recently proposed algorithm, by Xie et al. [33], for global network alignment fits very well to our problem with the following modifications.

Again consider the group in Figure 3. We take the clique network of the group and the network outside the group and place them apart as shown in box 3 of Figure 3. Next we consider three different networks. First network is the group clique network with adjacency matrix \mathbf{N}_g of size $(c \times c)$. Second network is the network outside this group with adjacency matrix \mathbf{N}_o of size $((n - c) \times (n - c))$. Third network is the inter-network which connects group member nodes to nodes external to group. Its adjacency matrix is \mathbf{X} of size $(c \times (n - c))$. Our aim is to learn the matrix \mathbf{R} whose each entry $\mathbf{R}(p, q)$ contains the affinity score for between the p^{th} group member and the q^{th} actor among the external actor nodes. We define a regularization framework, same as Xie et. al. [33], over the above three networks. The objective function for minimization is:

$$\min_{\mathbf{R}} \alpha \sum_{u, v, i, j} (\mathbf{N}_g \otimes \mathbf{N}_o)_{(i, u), (j, v)} (\mathbf{R}(i, u) - \mathbf{R}(j, v))^2 + (1 - \alpha) \sum_{i, u} (\mathbf{R}(i, u) - \mathbf{X}(i, u))^2 \quad (4)$$

where $\mathbf{N}_g \otimes \mathbf{N}_o$ is the Kronecker product of \mathbf{N}_g and \mathbf{N}_o . Each $(\mathbf{N}_g \otimes \mathbf{N}_o)_{(i, u), (j, v)}$ is 1 if $\mathbf{N}_g(i, j) = 1$ and $\mathbf{N}_o(u, v) = 1$, in other words i^{th} and j^{th} group members are linked (which should always be true as group is a clique) and u^{th} and v^{th} external actors are also linked, otherwise 0. The first term in the objective aligns group member i with external actor u and group member j with external actor v , if (i, j) are neighbors and (u, v) are also linked. This term therefore, enforces smoothness over \mathbf{R} . The second term is a regularization term that uses prior knowledge (like a and e are already connected in our example) stored in \mathbf{X} . $\alpha \in (0, 1]$ controls the trade-off between these two competing constraints. As proposed by Xie et. al., the most efficient method to minimize is by the following random-walks based recursive model:

$$\mathbf{R} = \alpha \mathbf{N}_g \mathbf{R} \mathbf{N}_o + (1 - \alpha) \mathbf{X} \quad (5)$$

The first term on the right hand side in the t^{th} recursive step becomes $\mathbf{N}_g^t \mathbf{R} \mathbf{N}_o^t$. This term mimics a random walk across the group network, inter network and the outer network. For $t = 1$ it represents cycles with group network path and outer network path length at most 1 as shown in Figure 3. In general in t^{th} step it captures cycles with path of length at most t in both clique and outer network. Notice the decay factor α penalizes the larger path length cycles recursively. For further algorithmic details we request to refer to Xie et. al.[33]. The exact algorithm used to learn \mathbf{R} for a given group is described in Appendix. Let \mathbf{R}_i represent the learned \mathbf{R} for the group g_i . Then the affinity of group g_i for j^{th} external actor is:

$$\mathbf{S}(i, j) = \frac{1}{c} \sum_{p=1}^c \mathbf{R}_i(p, q) \quad (6)$$

where q is the index in outer network \mathbf{N}_o for i^{th} actor in original NOA (\mathbf{A}). We therefore, learned affinity $\mathbf{S}(i, j)$, for i^{th} group and external actor, supervised using the existing connection of the group members to outer actor network.

C. Group Label Propagation Score (GLPS)

In the previous sections we developed methods that capture paths and cycles over the NOA but does not takes into account the NOG. In this section we develop label propagation based score which takes into account the hypergraph structure of the NOG. Intuitively, we start by giving some initial labels only to the members of a given i^{th} group g_i . These labels then diffuse by random-walks through the hypergraph structure of the NOG. Once the random-walks stabilize, the final label for each external vertex is treated as its affinity score for the given group. The final label at a given external actor vertex represents the chances a random walk originating from group member nodes might end up at this vertex. Therefore, modeling a network guided similarity between the group and the external actor.

To realize the above label diffusion as a hypergraph-based learning task. Let \mathbf{y} be the vector of initial labels to the vertices of the NOG hypergraph $N_g(V, G)$ with incidence matrix \mathbf{H} . For a “given” group g_i we have $\mathbf{y}(v) = 1$ if $v \in g_i$ else $\mathbf{y}(v) = 0$. We learn the final label vector \mathbf{f} . In order to take into account the hypergraph structure we want that the members (vertices) within “any” group (hyperedge) finally get same labels. Also we want the vertices of “given” group retain their initial labels. We capture these aims in the following cost minimization objective:

$$\min_{\mathbf{f}} \frac{1}{2} \sum_{g \in G} \sum_{u, v \in g} \frac{\mathbf{w}(g) \mathbf{H}(u, g) \mathbf{H}(v, g)}{\delta(g)} \left(\frac{\mathbf{f}(u)}{\sqrt{\mathbf{d}(u)}} - \frac{\mathbf{f}(v)}{\sqrt{\mathbf{d}(v)}} \right)^2 + \mu \|\mathbf{f} - \mathbf{y}\|^2 \quad (7)$$

where, \mathbf{w} is vector whose entries contain hyperedges weights, \mathbf{d} is vector containing vertex degrees such that for as vertex v : $\mathbf{d}(v) = \sum_{g \in G | v \in g} \mathbf{w}(g)$ and δ is vector containing hyperedge degrees such that for an edge g : $\delta(g) = \sum_{v \in V} \mathbf{H}(v, g)$. The first term is a smoothing term which makes sure that vertices within the same hyperedge have the same scores. So the more number of common hyperedges they are part of the more similar their score becomes. For example if two authors (vertex) have written several papers (hyperedge) together then they are more similar and should be assigned same scores. This term therefore, enforces the hypergraph structure while learning labels. The second term measures the difference between the given labels and the final vertex scores. The parameter μ then controls the degree of diffusion. In more compact matrix representation:

$$\min_{\mathbf{f}} \mathbf{f}^T \mathbf{L}_h \mathbf{f} + \mu \|\mathbf{f} - \mathbf{y}\|^2 \quad (8)$$

where,

$$\mathbf{L}_h = \mathbf{I} - \mathbf{D}_v^{-1/2} \mathbf{H} \mathbf{W} \mathbf{D}_e^{-1} \mathbf{H}^T \mathbf{D}_v^{-1/2} \quad (9)$$

is the normalized hypergraph laplacian [34], where \mathbf{D}_e and \mathbf{D}_v are diagonal matrices consisting of hyperedges and vertex degrees, with $(|G| \times |G|)$ and $(|V| \times |V|)$ sizes, respectively. \mathbf{W} is the $(|G| \times |G|)$ diagonal matrix containing weights of hyperedges. It is easy to show [34] that the solution to equation 8 is equivalent to solving the following linear system:

$$\mathbf{f}_i^* = (1 - \alpha)(\mathbf{I} - \alpha\theta)^{-1} \mathbf{y}, \quad (10)$$

where $\alpha = 1/(1 + \mu)$, $\theta = \mathbf{D}_v^{-1/2} \mathbf{H} \mathbf{W} \mathbf{D}_e^{-1} \mathbf{H}^T \mathbf{D}_v^{-1/2}$ and \mathbf{f}_i^* is the final label vector learned for the i^{th} group. The j^{th}

Table I. *Splits WITH FIXED BOUNDARY YEAR*

Boundary Yr	Split No.	Train	Test
1995	A.1	1992-95 (4 yrs)	1996-98 (3 yrs)
1995	A.2	1993-95 (3 yrs)	1996-98 (3 yrs)
1995	A.3	1993-95 (3 yrs)	1996-99(4 yrs)
2000	A.4	1997-00 (4 yrs)	2001-03 (3 yrs)
2000	A.5	1998-00 (3 yrs)	2001-03 (3 yrs)
2000	A.6	1998-00 (3 yrs)	2001-04 (4 yrs)
2005	A.7	2002-05 (4 yrs)	2006-08 (3 yrs)
2005	A.8	2003-05 (3 yrs)	2006-08 (3 yrs)
2005	A.9	2003-05 (3 yrs)	2006-09 (4 yrs)
2007	Main Split	2003-07 (5 yrs)	2008-10 (3 yrs)

entry in \mathbf{f}_i^* quantifies the affinity between of i^{th} group for the j^{th} actor. Therefore, we have:

$$\mathbf{S}(i, j) = \mathbf{f}_i^*(j). \quad (11)$$

Since, our aim is to predict a score for each of the IGs g_i^{in} (or SGs g_i^{sa}) for all $i \in \{1, \dots, m\}$. We treat the affinity of group g_i for j^{th} actor ($\mathbf{S}(i, j)$) as the score which reflects the possibility of IG $g_i^a \in g_i^{in}$ (actor a is v_j) being formed in future. In summary, the prediction score for g_i^a (where a is v_j) is taken as $\mathbf{S}(i, j)$. In fact we also assign $\mathbf{S}(i, j)$ as the score for SG s_i^a , same as that for IG g_i^a . Note that one can build more complicated scores, eg: weighted by the size of subgroup, etc. But for this study we restrict ourselves to this simplified scenario. In experimentation section we shall further discuss how to get the ranked list of most likely future groups using these prediction scores.

V. EXPERIMENTAL ANALYSIS

In the following the first section describes the dataset statistics. Second section is about evaluation metrics used followed by experiments and analysis in third section.

A. Dataset and Statistics

In this paper we have used the popular DBLP dataset (publicly available at [35]) and extracted all the publication from years 1930-2011 for top 10 venues, as listed in <http://academic.research.microsoft.com/>, each for 22 different sub-fields of computer science (total 220 top venues). Here we analyze some interesting statistical properties of this dataset which has motivated this research. Each publication is written by a group of authors and the same group can write multiple publications as well. For both, statistics and experiments we have divided the dataset into various training and test periods (*splits*) as shown in the Tables I. In Table I, each row is a split with a fixed end year of training set (boundary year). Table II contains the statistics for (*sub*)*incremental* groups present in “test” period of different splits. In Table II we refer an actor or group as **old** if it has been observed in training else we use **new**. Note in case a group has written multiple papers in a train or test period it is counted only once. We observe that on an average around 20% of the groups formed in test period contain no new authors (old actor groups (OAG)). Out of these upto approx 20% are *incremental* groups (IGs) formed by IA process. Moreover, around 80% of these IGs are new and never observed in training. Also, we notice that approx 70% of the groups (with no new author (OAG)) are *subincremental* groups (SGs) formed by SA. All these percentages are highlighted in the Table II. These subtle observations indicate that IA and SA processes are responsible for a large portion of the

Table II. INCREMENTAL STATISTICS OF TESTING PERIODS FOR THE *Splits* IN TABLE I

Split No.	# New Actor Groups (NAG)	# Old Actor Groups (OAG)	# Total Groups	% New Actor Groups	% Old Actor Groups	# Old IGs	# New IGs	Total IGs	% (of OAG) New IGs	% (of OAG) Old IGs	% (of OAG) Total IGs	Total % (of IGs) New IGs	Total % (of IGs) Old IGs
A.1	7863	2343	10206	77.043	22.95	113	386	499	16.47	4.82	21.29	77.36	22.64
A.2	8004	2202	10206	78.42	21.57	81	350	431	15.89	3.68	19.57	81.21	18.79
A.3	11665	2623	14288	81.64	18.36	91	416	507	15.86	3.47	19.33	82.05	17.94
A.4	14308	3629	17937	79.77	20.23	166	550	716	15.16	4.57	19.73	76.81	23.18
A.5	14543	3394	17937	81.08	18.92	139	470	609	13.85	4.095	17.94	77.17	22.82
A.6	20331	3872	24203	84.00	15.99	146	543	689	14.02	3.77	17.79	78.81	21.20
A.7	25081	6351	31432	79.79	20.20	285	936	1221	14.74	4.49	19.23	76.65	23.34
A.8	25482	5950	31432	81.30	18.98	230	796	1026	13.38	3.86	17.24	77.58	22.41
A.9	34747	6827	41474	83.78	16.46	244	897	1141	13.14	3.57	16.71	78.61	21.38
Avg.				80.76	19.30				14.72	4.04	18.76	78.47	21.52
Main Split (IA)	25149	7624	32773	23.26	76.74	308	1085	1393	14.23	4.04	18.27	77.89	21.11
Main Split (SA)	25149	7624	32773	23.26	76.73	1503	3825	5328	50.17	19.71	69.88	71.79	28.20

OAGs formed in future. Therefore, modeling these processes is an important step towards higher order link prediction. Constrained by space limit we have not shown SG statistics (except for the **main split** in Table II). Also the number of actors per split roughly range from 30K to 100K ($> 100K$ in **main split**) and there are 10K to 30K groups in “training” period across various splits.

B. Evaluation Methodology and Experimental Setup

In this section we describe two different kinds of metrics, **global** and **per-group**. The performance of the proposed approaches is evaluated using the training (2004-07) and testing period (2008-10) of **main split** in Table I. The statistics of main split are shown at bottom of Table II. For the groups in the training period, each method: GKS, BRWS and GLPS, was run to output the scores for all IGs (g_i^{in}) and all SGs (g_i^{sa}) for each group ($i \in \{1, \dots, m\}$). Now consider the whole set containing all the IGs for all groups. As there might be many repeating IGs we shall only consider unique IGs while taking the maximum score among all the repeating IGs. We sort this unique set of IGs by their scores and get the highest scoring top- N_{top} IGs. We do the exact same thing for SG case and get top- N_{top} SGs. Out of these top- N_{top} groups (IG or SG) the performance over test set (2008-10) is compared using the following metrics:

$$\text{Precision}@N_{top} \text{ (IA)} = \frac{\text{Number of groups correctly predicted using IA process from top-}N_{top} \text{ list}}{N_{top}} \quad (12)$$

$$\text{Recall}@N_{top} \text{ (IA)} = \frac{\text{Number of collaborations correctly predicted using IA process from top-}N_{top} \text{ list}}{\text{\# of actual IA generated groups}} \quad (13)$$

$$\text{Precision}@N_{top} \text{ (SA)} = \frac{\text{Number of groups correctly predicted using SA process from top-}N_{top} \text{ list}}{N_{top}} \quad (14)$$

$$\text{Recall}@N_{top} \text{ (SA)} = \frac{\text{Number of collaborations correctly predicted using SA process from top-}N_{top} \text{ list}}{\text{\# of actual SA generated groups}} \quad (15)$$

The above **global** metrics (equations 12 to 15) capture overall predictions across all groups. But often times we are more interested in understanding the future of a single group and how it will evolve in future. For this case we simply sort

IGs g_i^{in} (or SGs g_i^{sa}) by their scores in descending order, to get the top- N_{top}^g IGs (or SGs) for each i^{th} group. Therefore, for this case we define the following metrics:

$$i^{th} \text{GroupPrecision}@N_{top}^g \text{ (IA)} = \frac{\text{Number of groups correctly predicted by IA of } i^{th} \text{ group from top-}N_{top}^g \text{ list}}{N_{top}^g} \quad (16)$$

$$i^{th} \text{GroupRecall}@N_{top}^g \text{ (IA)} = \frac{\text{Number of groups correctly predicted by IA of } i^{th} \text{ group from top-}N_{top}^g \text{ list}}{\text{\# of actual IA generated groups from the } i^{th} \text{ group}} \quad (17)$$

for each i^{th} group and take average of these metrics to derive the following average metrics:

$$\text{AvgPrecision}@N_{top}^g \text{ (IA)} = \frac{\text{Sum of } i^{th} \text{GroupPrecision}@N_{top}^g \text{ (IA) for all groups in training set}}{\text{Total \# of groups in training set}} \quad (18)$$

$$\text{AvgRecall}@N_{top}^g \text{ (IA)} = \frac{\text{Sum of } i^{th} \text{GroupRecall}@N_{top}^g \text{ (IA) for all groups in training set}}{\text{Total \# of groups in training set}} \quad (19)$$

We refer to the metrics in equations 18 to 19 as the **per-group** metrics. Metrics analogous to equation 16 to 19: $i^{th} \text{GroupPrecision}@N_{top}^g \text{ (SA)}$, $i^{th} \text{GroupRecall}@N_{top}^g \text{ (SA)}$, $\text{AvgPrecision}@N_{top}^g \text{ (SA)}$ and $\text{AvgRecall}@N_{top}^g \text{ (SA)}$, are defined for the SA case as well.

All the three methods GKS, BRWS and GLPS were implemented in MATLAB. For GKS, BRWS and GLPS the parameters $\beta = \{0.1, 0.3, 0.5, 0.7, 0.9\}$, $\alpha = \{0.1, 0.2, 0.4, 0.5, 0.6, 0.8, 0.9\}$, and $\mu = \{0.1, 0.3, 0.5, 0.7, 0.9\}$, respectively, were tested. Using 10-fold cross-validation the following best values were observed: $\{\beta = 0.5, \alpha = 0.6, \mu = 0.1\}$ for global metrics and $\{\beta = 0.5, \alpha = 0.6, \mu = 0.5\}$ for per-group metrics. All the hypergraphs and graphs considered in any of the methods are all unweighted and $l \leq 4$ is only considered in GKS. Experiments were run using Intel Core i7 (2.8 GHz) CPU with 4 GB RAM.

C. Results and Discussion

In this section we discuss the results obtained using both, the **global** and the **per-group** metrics, for the **main split**.

Note that we omit the results from other splits due to space constraints and moreover, they showed results very similar to the main split. Our **GKS** method simply extends Katz (1953) [22] (which is among the most successful DLP methods [31]). We therefore, consider **GKS** as a strong baseline for our evaluation. We would also like to mention that we had explored a number of matrix factorization (MF) based DLP methods [30] like MF, Non-Negative MF, SVD, Tri-MF and their variants with (or without) network regularization and sparsity constraints. But all of them performed trivially in comparison to our methods, so we don't include them.

We now discuss the results for **per-group** and **global** metrics. Notice that in per-group scenario we inspect each group individually and compare its affinity for different actors outside it. Therefore, in this case our comparison is between various "actors" to ascertain which actors are more likely to join a given group. Whereas in case of global metrics we try to compare scores of different "groups" (IG or SG) and tell which are more likely to occur in future. Keeping this in mind let us first consider the per-group metrics results in Table III for $N_{top}^g = 100$. We observe that both **GLPS** and **BRWS** outperforms **GKS** (baseline) consistently for both IA and SA cases. Notice that both **GLPS** and **BRWS** are semi-supervised algorithms, with former supervised by the hypergraph structure and the later learns possible cyclic connections from the existing connections between a group with outside actors. In contrast, **GKS** which simply works on path enumeration based similarity calculation and lacks supervision performs bad. The good performance of **GLPS** and **BRWS** suggests that both: (1) hypergraph structure (i.e. how groups as composite entities are connected to each other and therefore, also to the groups in which an external actor participates?) (2) cycles passing through group and an external actor node (i.e. which all group members an external actor knows and through which communication cycles?). Similar, trend is observed in results for global metrics shown in Table IV for $N_{tot} = 10000$. Again, **GLPS** and **BRWS** perform better than **GKS**. However, in this case **GLPS** does better than **BRWS** in SA scenario. A possible explanation lies in the fact that **GLPS** keeps the group as well as subgroup structure intact using the hypergraph model. As an example if we have observed a group P containing actors $\{x, y, z, w\}$ and also its subgroup Q with actors $\{x, y, z\}$. While evaluating group P, **BRWS** will not consider the existence of the P's subgroup Q as it models the NOA not NOG. Whereas, **GLPS** models NOG and keep both group as well as subgroup information intact in the hypergraph laplacian of the NOG. This attribute of **GLPS**, to capture subgroups within other groups, helps it to outperform in SA, where this distinction between groups and its subgroups is quiet critical.

Another point to mention is the low values for global metrics and precision in case of per group metrics. The reason for this is due to the inherent difficulty of the problem at hand. The number of groups formed by (sub)incremental accretion processes (positives occurrence), though a considerable portion of groups in testing period, are much smaller than the total possibilities. Given a group of size r and n as the total number of actors ($> 100K$ in main split) in the network. There are $PIA = (n - r)$ and $PSA = \{(2^r - 2) \times (n - r)\}$, number of IGs and SGs possible respectively, from the given group. The large number of actors $n (\approx 10^5)$ makes $PIA (\approx 10^5)$ large and $PSA (\approx 10^6)$ (restricting to group size $r_{max} \leq 6$),

Table III. MAIN SPLIT **PER-GROUP** METRICS RESULTS

	GKS	GLPS	BRWS
AvgPrecision@100 (IA)	0.0210	0.0349	0.0355
AvgRecall@100 (IA)	0.3176	0.6034	0.6050
AvgPrecision@100 (SA)	0.0198	0.0266	0.0271
AvgRecall@100 (SA)	0.2616	0.5149	0.5135

Table IV. MAIN SPLIT **GLOBAL** METRICS RESULTS

	GKS	GLPS	BRWS
Precision@10000 (IA)	0.0020	0.0075	0.0134
Recall@10000 (IA)	0.0144	0.0538	0.0962
Precision@10000 (SA)	0.0052	0.0666	0.0327
Recall@10000 (SA)	0.0098	0.125	0.0614

even much larger, in worst case. Even though, $PIA < PSA$ but the number of IGs formed in test period on an average is much lesser (20%) as compared to (70%) of SGs (V-A). Due to this, in case of per group metric, chances of finding a positive occurrences within $N_{tot}^g = 100$ groups out of PIA or PSA possibilities, is quiet challenging task. This explains low precision in per group scenario. Global metric scenario is even worse. Assume m (around $30K$ in main split) groups and let us say, for approximation purpose, all are of size r . Then there are $GIA = \{m \times PIA\} \approx 10^9$ and $GSA = \{m \times PSA\} \approx 10^{10}$ number of total IGs and SGs possible (assuming $r_{max} \leq 6$, $m \approx 10^4$ and $n \approx 10^5$). In case of global metric finding all positive occurrences within just $N_{tot} = 10^4$ groups out of the huge GIA and GSA possibilities explains the low global precision scores. (Note above is a rough worst approximation in which we have restricted $r_{max} \leq 6$ for illustration. In depth discussion will involve actual group cardinality distribution which we leave due to space limitations.) However, there are a limited number of IA and SA generated groups. We can therefore, hope to cover a significant portion of them within top ranked groups. This makes recall more important measure for us and it attains significantly high values as compared to precision at least for the per group metrics. However, in case of global metrics the number of possibilities are huge, resulting in low values for recall as well.

VI. CONCLUSIONS

In this work we have addressed the problem of evolution of Small Groups while highlighting its differences with the general problem of community evolution. We found statistically that two *group accretion* processes are behind the formation of a large percentage of future groups given past history of group collaborations. We have built different models that capture these two processes while being motivated from different theories from social science. We treat the problem of future group prediction as a higher-order link prediction task and have developed three topology based methods. Extensive experiments carried out using DBLP dataset show that our methods give good results while predicting future groups.

Acknowledgement This work is supported by NSF Award IIS-1422802 and BBN Contract W911NF-09-2-0053.

VII. APPENDIX

Bi-random Walk Algorithm In their paper Xie et. al. [33] have provided various versions of Bi-random Walk algorithm. In our work we use the sequential version **BiRW_seq**. Input for the algorithm are: the group network adjacency matrix N_g , the inter network matrix I , outer network matrix N_o ,

the decay parameter α , and l_g and l_o are the maximum path length allowed while generating cycles in the group and outer networks respectively. While generating cycles **BiRW_seq** does a random-walk on the group network followed by a random-walk on outer network sequentially in each step. In line 3 the algorithm takes a random-walk over the group clique if the length of the path in the group network is still less than l_g . A second step is taken on the outer network if path on the outer network covered till now is less than l_o (line 6). In our implementation both l_g and l_o are taken as 4. On reaching maximum path lengths on both networks **R** is returned.

Algorithm 1 BiRW_seq($N_g, I, N_o, \alpha, l_g, l_o$)

```

1:  $R^0 = \frac{I}{\text{sum}(I)}$ 
2: for all  $t = 1$  to  $\max\{l_g, l_r\}$  do
3:   if  $t \leq l_g$  then
4:      $R^{t_{group}} = \alpha N_g R^{t-1} + (1 - \alpha)I$ 
5:   end if
6:   if  $t \leq l_o$  then
7:      $R^t = \alpha R^{t_{group}} N_o + (1 - \alpha)I$ 
8:   end if
9: end for
10: return (R)

```

REFERENCES

- [1] S. Poole, M. S. Poole, and A. B. Hollingshead, *Theories of small groups: Interdisciplinary perspectives*. Sage Publications, 2004.
- [2] G. Cheney, L. T. Christensen, T. E. Zorn Jr, and S. Ganesh, *Organizational communication in an age of globalization: Issues, reflections, practices*. Waveland Press, 2010.
- [3] W. F. Boh, Y. Ren, S. Kiesler, and R. Bussjaeger, "Expertise and collaboration in the geographically dispersed organization," *Organization Science*, vol. 18, no. 4, pp. 595–612, 2007.
- [4] C. S. Wagner, J. D. Roessner, K. Bobb, J. T. Klein, K. W. Boyack, J. Keyton, I. Rafols, and K. Börner, "Approaches to understanding and measuring interdisciplinary scientific research (idr): A review of the literature," *Journal of Informetrics*, vol. 5, no. 1, pp. 14–26, 2011.
- [5] A. Lungeanu, Y. Huang, and N. S. Contractor, "Understanding the assembly of interdisciplinary teams and its impact on performance," *Journal of informetrics*, vol. 8, no. 1, pp. 59–70, 2014.
- [6] J. Hahn, J. Y. Moon, and C. Zhang, "Emergence of new project teams from open source software developer networks: Impact of prior collaboration ties," *Information Systems Research*, vol. 19, no. 3, pp. 369–391, 2008.
- [7] N. Contractor, "Some assembly required: leveraging web science to understand and enable team assembly," *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, vol. 371, no. 1987, p. 20120385, 2013.
- [8] A. Patil, J. Liu, and J. Gao, "Predicting group stability in online social networks," in *Proceedings of the 22nd international conference on World Wide Web*. International World Wide Web Conferences Steering Committee, 2013, pp. 1021–1030.
- [9] A. Patil, J. Liu, B. Price, H. Sharara, and O. Brdiczka, "Modeling destructive group dynamics in online gaming communities," in *Proc. of the 6th Int. AAAI Conf. on Weblogs and Social Media*, vol. 2012, 2012.
- [10] H. Sharara, L. Singh, L. Getoor, and J. Mann, "Stability vs. diversity: Understanding the dynamics of actors in time-varying affiliation networks," in *Social Informatics (SocialInformatics), 2012 International Conference on*. IEEE, 2012, pp. 1–6.
- [11] A. R. Kang, J. Park, and H. K. Kim, "Loyalty or profit? early evolutionary dynamics of online game groups," in *Network and Systems Support for Games (NetGames), 2013 12th Annual Workshop on*. IEEE, 2013, pp. 1–6.
- [12] N. F. Johnson, C. Xu, Z. Zhao, N. Ducheneaut, N. Yee, G. Tita, and P. M. Hui, "Human group formation in online guilds and offline gangs driven by a common team dynamic," *Physical Review E*, vol. 79, no. 6, p. 066117, 2009.
- [13] M. A. Ahmad, Z. Borbora, C. Shen, J. Srivastava, and D. Williams, *Guild play in mMOGs: Rethinking common group dynamics models*. Springer, 2011.
- [14] H. Alvari, K. Lakkaraju, G. Sukthankar, and J. Whetzel, "Predicting guild membership in massively multiplayer online games," in *Social Computing, Behavioral-Cultural Modeling and Prediction*. Springer, 2014, pp. 215–222.
- [15] C.-H. Chen, C.-T. Sun, and J. Hsieh, "Player guild dynamics and evolution in massively multiplayer online games," *CyberPsychology & Behavior*, vol. 11, no. 3, pp. 293–301, 2008.
- [16] A. Patil, J. Liu, J. Shen, O. Brdiczka, J. Gao, and J. Hanley, "Modeling attrition in organizations from email communication," in *Social Computing (SocialCom), 2013 International Conference on*. IEEE, 2013, pp. 331–338.
- [17] H. Sharara, L. Singh, L. Getoor, and J. Mann, "The dynamics of actor loyalty to groups in affiliation networks," in *Social Network Analysis and Mining, 2009. ASONAM'09. International Conference on Advances in*. IEEE, 2009, pp. 101–106.
- [18] M. Spiliopoulou, "Evolution in social networks: A survey," in *Social network data analytics*. Springer, 2011, pp. 149–175.
- [19] S. A. Beebe and J. T. Masterson, "Communication in small groups: principles and practice," 2009.
- [20] L. L. Putnam and C. Stohl, "Bona fide groups," *Research Methods for Studying Groups and Teams: A Guide to Approaches, Tools, and Technologies*, p. 211, 2012.
- [21] P. R. Monge and N. S. Contractor, *Theories of communication networks*. Oxford University Press, 2003.
- [22] L. Katz, "A new status index derived from sociometric analysis," *Psychometrika*, vol. 18, no. 1, pp. 39–43, 1953.
- [23] C. Berge and E. Minieka, *Graphs and hypergraphs*. North-Holland publishing company Amsterdam, 1973, vol. 7.
- [24] C. Aggarwal and K. Subbian, "Evolutionary network analysis: A survey," *ACM Computing Surveys (CSUR)*, vol. 47, no. 1, p. 10, 2014.
- [25] H. Alvari, S. Hashemi, and A. Hamzeh, "Detecting overlapping communities in social networks by game theory and structural equivalence concept," in *Artificial Intelligence and Computational Intelligence*. Springer, 2011, pp. 620–630.
- [26] H. Alvari, A. Hajibagheri, and G. Sukthankar, "Community detection in dynamic social networks: A game-theoretic approach."
- [27] T. Chung, J. Han, D. Choi, T. T. Kwon, H. K. Kim, and Y. Choi, "Unveiling group characteristics in online social games: A socio-economic analysis," in *Proceedings of the 23rd International Conference on World Wide Web*, ser. WWW '14. International World Wide Web Conferences Steering Committee, 2014.
- [28] H. Oh, M.-H. Chung, and G. Labianca, "Group social capital and group effectiveness: The role of informal socializing ties," *Academy of Management Journal*, vol. 47, no. 6, pp. 860–875, 2004.
- [29] L. Lü and T. Zhou, "Link prediction in complex networks: A survey," *Physica A: Statistical Mechanics and its Applications*, vol. 390, no. 6, pp. 1150–1170, 2011.
- [30] M. Al Hasan and M. J. Zaki, "A survey of link prediction in social networks," in *Social network data analytics*. Springer, 2011, pp. 243–275.
- [31] D. Liben-Nowell and J. Kleinberg, "The link-prediction problem for social networks," *Journal of the American society for information science and technology*, vol. 58, no. 7, pp. 1019–1031, 2007.
- [32] J. Flannick, *Algorithms for biological network alignment*. ProQuest, 2008.
- [33] M. Xie, T. Hwang, and R. Kuang, "Prioritizing disease genes by bi-random walk," in *Advances in Knowledge Discovery and Data Mining*. Springer, 2012, pp. 292–303.
- [34] D. Zhou, J. Huang, and B. Schölkopf, "Learning with hypergraphs: Clustering, classification, and embedding," in *Advances in neural information processing systems*, 2006, pp. 1601–1608.
- [35] J. Tang, D. Zhang, and L. Yao, "Social network extraction of academic researchers," in *Data Mining, 2007. ICDM 2007. Seventh IEEE International Conference on*. IEEE, 2007, pp. 292–301.